Error threshold in finite populations

D. Alves and J. F. Fontanari

Instituto de Física de São Carlos

Universidade de São Paulo

Caixa Postal 369

13560-970 São Carlos SP

Brazil

Abstract

A simple analytical framework to study the molecular quasispecies evolution of finite populations is proposed, in which the population is assumed to be a random combination of the constituent molecules in each generation, i.e., linkage disequilibrium at the population level is neglected. In particular, for the single-sharp-peak replication landscape we investigate the dependence of the error threshold on the population size and find that the replication accuracy at the threshold increases linearly with the reciprocal of the population size for sufficiently large populations. Furthermore, in the deterministic limit our formulation yields the exact steady-state of the quasispecies model, indicating then that the population composition is a random combination of the molecules.

87.10.+e, 64.60.Cn

I. INTRODUCTION

An important issue in the investigation of the dynamics of competing self-reproducing macromolecules, whose paradigm is Eigen's quasispecies model [1], is the effect of the finite size of the population on the error threshold phenomenon that limits the length of the molecules [2]. The quasispecies model was originally formulated as a deterministic kinetic theory described by a set of ordinary differential equations for the concentrations of the different types of molecules that compose the population. Such formulation, however, is valid only in the limit where the total number of molecules N goes to infinity. More pointedly, in this model a molecule is represented by a string of ν digits $(s_1, s_2, \ldots, s_\nu)$, with the variables s_α allowed to take on κ different values, each of which representing a different type of monomer used to build the molecule. For sake of simplicity, in this paper we will consider only binary strings, i.e., $s_\alpha = 0, 1$. The concentrations x_i of molecules of type $i = 1, 2, \ldots, 2^\nu$ evolve in time according to the following differential equations [1,2]

$$\frac{dx_i}{dt} = \sum_j W_{ij} x_j - \left[D_i + \Phi\left(t\right)\right] x_i \ , \tag{1}$$

where the constants D_i stand for the death probability of molecules of type i, and $\Phi(t)$ is a dilution flux that keeps the total concentration constant. This flux introduces a nonlinearity in (1), and is determined by the condition $\sum_i dx_i/dt = 0$. The elements of the replication matrix W_{ij} depend on the replication rate or fitness A_i of the molecules of type i as well as on the Hamming distance $d\left(i, j\right)$ between strings i and j. They are given by

$$W_{ii} = A_i \, q^\nu \tag{2}$$

and

$$W_{ij} = A_i \, q^{\nu - d(i,j)} \left(1 - q\right)^{d(i,j)} \qquad i \neq j, \tag{3}$$

where $0 \leq q \leq 1$ is the single-digit replication accuracy, which is assumed to be the same for all digits. Henceforth we will set $D_i = 0$ for all i. The quasispecies concept is illustrated more neatly for the single-sharp-peak replication landscape, in which we ascribe the replication

rate $a > 1$ to the so-called master string $(1, 1, \ldots, 1)$, and the replication rate 1 to the remaining strings. In this context, the parameter a is termed selective advantage of the master string. As the error rate $1 - q$ increases, two distinct regimes are observed in the population composition: the *quasispecies* regime characterized by the master string and its close neighbors, and the *uniform* regime where the 2^ν strings appear in the same proportion. The transition between these regimes takes place at the error threshold $1 - q_t$, whose value depends on the parameters ν and a [1,2]. A genuine thermodynamic order-disorder phase transition occurs in the limit $\nu \to \infty$ only [3–5]. We must note, however, that standard statistical mechanics tools developed to study the surface equilibrium properties of lattice systems can be used to investigate the finite ν case as well [3,4]. Moreover, the complete analytical solution of the single-sharp-peak replication landscape has been found recently by mapping the stationary solution of the kinetic equations (1) into a polymer localization problem [5,6].

Closely related to our approach to the quasispecies evolution of finite populations is the population genetics formulation of the deterministic quasispecies model proposed recently [7]. In that formulation it is assumed that the molecules are characterized solely by the number of monomers 1 they have, regardless of the particular positions of these monomers inside the molecules. Hence there are only $\nu+1$ different types of molecules which are labeled by the integer $P = 0, 1, \ldots, \nu$. This assumption is not so far-fetched since the feature that distinguishes the molecules is their replication rates A_i, which in most analyses have been chosen to depend on P only, i.e., $A_i = A_P$ [2]. Furthermore, denoting the frequency of monomers 1 in generation t by p_t, it is assumed that the molecule frequencies $\Pi_P(t)$ are given by the binomial distribution

$$\Pi_P(t) = \binom{\nu}{P} (p_t)^P \, (1 - p_t)^{\nu - P} \tag{4}$$

for $P = 0, 1, \ldots, \nu$. Thus, in each generation the monomers are sampled with replacement from a pool containing monomers 1 and 0 in the proportions p_t and $1 - p_t$, respectively. This amounts to neglecting linkage disequilibrium, i.e., in each generation the molecule frequen-

cies are random combinations of the constituent monomers [8]. With the two assumptions presented above a simple recursion relation for the monomer frequency p_t can be readily derived [7].

To take into account the effect of finite N, the deterministic kinetic formulation must be replaced by a stochastic formulation based on a master equation for the probability distribution of the number of different types of molecules in the population [9,10]. However, the extreme approximations used to derive results from that master equation or from related Langevin equations [11,12] have hindered the analysis of the error threshold for finite populations. An alternative approach to study stochastic chemical reaction networks is the algorithm proposed by Gillespie [13], which has been successfully employed to simulate numerically the quasispecies model, providing thus a base line for analytical investigations [14]. The goal of this work is to propose an analytical framework to investigate the quasispecies evolution of finite populations. More specifically, we will focus on the evolution of the molecule frequencies $\Pi_P(t)$ for $P = 0, \ldots, \nu$ and, since for finite N these frequencies are random variables, we will derive a recursion relation for the average values $\overline{\Pi}_P(t)$. Although we will concentrate mainly on the dependence of the error threshold on the population size N, the formalism presented in the sequel can be applied to study a variety of fascinating phenomena related to the finitude of the population, such as mutational meltdown [15] and punctuated equilibria or stasis [12,16], to mention only a few. Moreover, since modern theories of integration of information in pre-biotic systems involve the compartmentation of a small number of molecules (typically 10 to 100) [17], the understanding of the effects of the error propagation in finite populations has become an important issue to the theories of the origin of life.

II. THE MODEL

In each generation the population is described by the vector $\mathbf{n} = (n_0, \ldots, n_\nu)$ where n_P is the number of molecules of type P, so that $\sum_P n_P = N$. Similarly to the deterministic

case [7], we have to resort to a simplifying assumption to relate the molecule frequencies Π_P to the vector $\mathbf{n}$. In particular, in generation t we consider a molecule pool containing the different molecule types in the proportions Π_P, so that $\mathbf{n}$ is distributed by the multinomial distribution

$$\mathcal{P}(\mathbf{n}) = \frac{N!}{n_0!\, n_1! \ldots n_\nu!} \left[\Pi_0(t)\right]^{n_0} \left[\Pi_1(t)\right]^{n_1} \ldots \left[\Pi_\nu(t)\right]^{n_\nu}. \tag{5}$$

Hence in each generation the molecules are sampled with replacement from the molecule pool. In this sense, in each generation the population is a random combination of the constituent molecules, which amounts to neglecting linkage disequilibrium at the population level. More pointedly, the population composed of the offspring of the molecules present in the generation $t-1$ is destroyed and its molecule frequency $\Pi_P(t)$ used to create an entire new population according to (5). Although this procedure destroys the correlations between the molecules, it does not cause any significant loss of genetic information since the fitness of the molecules depend only on the number of monomers 1 they have, which, in the average, is not affected by the procedure.

The changes in the population composition $\mathbf{n}$ are due to the driving of natural selection, modeled by the replication rate A_P, and to mutations, modeled by the error rate per digit $1-q$. Following the prescription used in the implementation of the standard genetic algorithm [18], we consider first the effect of natural selection and then the effect of mutations. As usual we assume that the number of offspring that each molecule contributes to the new generation is proportional to its relative replication rate which, for molecules of type P, is defined by

$$W_P(\mathbf{n}) = \frac{n_P A_P}{\sum_R n_R A_R}. \tag{6}$$

Thus the population composition after selection is described by the random vector $\mathbf{n}' = (n'_0, \ldots, n'_\nu)$ which is distributed according to the conditional probability distribution

$$\mathcal{P}_s(\mathbf{n}' \mid \mathbf{n}) = \frac{N!}{n'_0!\, n'_1! \ldots n'_\nu!} \left[W_0(\mathbf{n})\right]^{n'_0} \left[W_1(\mathbf{n})\right]^{n'_1} \ldots \left[W_\nu(\mathbf{n})\right]^{n'_\nu}. \tag{7}$$

Next we consider the changes in $\mathbf{n}'$ due to mutations. After mutation, the population is described by $\mathbf{n}'' = (n''_0, \ldots, n''_\nu)$ whose components are written as

$$n''_P = \sum_{R=0}^{\nu} n''_{PR}, \tag{8}$$

where the integer n''_{PR} stands for the number of molecules of type R that have mutated to a molecule of type P. Clearly, $n'_R = \sum_P n''_{PR}$. We note that the probability of mutation from a molecule of type R to a molecule of type P is given by

$$M_{PR} = \sum_{Q=Q_l}^{Q_u} \binom{R}{Q} \binom{\nu - R}{P - Q} q^{\nu - P - R + 2Q} \, (1-q)^{P+R-2Q}, \tag{9}$$

where $Q_l = \max(0, P + R - \nu)$ and $Q_u = \min(P, R)$. The population is more conveniently described by the set $\{n''_{PR}\}$ rather than by $\mathbf{n}''$. In fact, given n'_R the conditional probability distribution of $\{n''_{PR}\}$ is again a multinomial

$$\mathcal{P}_m \left(n''_{0R}, n''_{1R}, \ldots, n''_{\nu R} \mid n'_R\right) = \frac{n'_R!}{n''_{0R}! \, n''_{1R}! \ldots n''_{\nu R}!} \, M_{0R}^{n''_{0R}} \, M_{1R}^{n''_{1R}} \ldots M_{\nu R}^{n''_{\nu R}} \tag{10}$$

for $R = 0, \ldots, \nu$. In this framework the frequency of molecules of type P in the next generation $\Pi_P(t+1)$ is given simply by $\frac{1}{N} \sum_R n''_{PR}$. This frequency is used to generate the new population of N molecules of length ν according to the distribution (5). The procedure is then repeated again.

We have run simulations for the single-sharp-peak replication landscape using the procedure described above, which neglects linkage disequilibrium at the population level, as well as the standard genetic algorithm [18], in which the correlations between consecutive generations are maintained. We have focused on the effect of the error rate $1 - q$ on the normalized mean Hamming distance d between the master string and the whole population in the stationary regime. This quantity is given by the fraction of monomers 0 in the entire population, i.e., $d = \frac{1}{\nu} \sum_P (\nu - P) \Pi_P$. In Figs. (1) and (2) we present the results of the simulations for d and its standard deviation σ, respectively, as functions of the error rate $1 - q$. The initial population is set with $\Pi_\nu = 1$ and $\Pi_P = 0$ for $P \neq \nu$, and it is left to evolve for $2 \; 10^3$ generations. No significant differences were found for longer runs or for different

choices of the initial molecular frequencies. Each data point involves two kinds of average: for each run we average over the mean Hamming distance in the last 100 generations; this value is then averaged over 200 runs. We note that even if the populations are identical in the initial generation, the random character of the transitions $\mathbf{n} \to \mathbf{n}' \to \mathbf{n}''$ will make them distinct in the next generation. It is clear from these results that, as N increases, the quantitative effects of assumption (5) become less significant. Moreover, the dependence of d and σ on the error rate is qualitatively the same for both algorithms.

III. RECURSION EQUATIONS

To derive an analytical recursion relation for the average molecular frequencies $\overline{\Pi}_P(t)$ we consider the following approximate procedure, akin to the annealed approximation of the statistical mechanics of disordered systems, which facilitates greatly the analysis: instead of averaging over the populations only after the stationary regime is reached, we perform this average in each generation. The result obtained $\overline{\Pi}_P(t)$ is then used to build the new populations. Of course, in doing so we neglect the fluctuations of $\Pi_P(t)$ for the different runs. Within this framework the average frequency of molecules of type P in generation $t+1$ is written as

$$\overline{\Pi}_P(t+1) = \frac{1}{N} \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \sum_{R} n''_{PR}\, \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}\left(\mathbf{n}\right). \tag{11}$$

Using

$$\sum_{\{n''_{PR}\}} n''_{PR}\, \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) = M_{PR} n'_R \tag{12}$$

and

$$\sum_{n'_R} n'_R \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) = N W_R\left(\mathbf{n}\right) \tag{13}$$

we rewrite (11) as

$$\overline{\Pi}_P(t+1) = \sum_{\mathbf{n}} \sum_{R} M_{PR} W_R\left(\mathbf{n}\right) \mathcal{P}\left(\mathbf{n}\right). \tag{14}$$

Noting that $\sum_P M_{PR} = 1$ and $\sum_R W_R(\mathbf{n}) = 1$, we can easily verify that the normalization condition $\sum_P \overline{\Pi}_P(t+1) = 1$ is satisfied. To proceed further we must specify the replication rate A_P. In the case of the single-sharp-peak replication landscape, i.e., $A_\nu = a$ and $A_P = 1$ for $P \neq \nu$, the summations over $n_0, \ldots, n_{\nu-1}$ can be readily carried out. The final result is

$$\overline{\Pi}_P(t+1) = M_{P\nu}\left[\overline{\Pi}_\nu(t)\right]^N + \sum_{n=0}^{N-1} B_n \frac{\sum_{R=0}^{\nu-1} \overline{\Pi}_R(t)\left[M_{PR} + a\frac{r}{1-r}M_{P\nu}\right]}{1 + r\,(a-1)} \tag{15}$$

for $P = 0, \ldots, \nu$. Here

$$B_n = \binom{N-1}{n}\left[\overline{\Pi}_\nu(t)\right]^n\left[1 - \overline{\Pi}_\nu(t)\right]^{N-1-n}, \tag{16}$$

and $r = n/N$. Thus, given the initial average molecular frequencies $\overline{\Pi}_P(t=0)$ for $P = 0, \ldots, \nu$, equations (15) are iterated till the stationary regime is reached.

Before we proceed on the analysis of the stationary solutions of the recursion equations (15), some comments regarding the definition of the error threshold are in order. A popular definition of error threshold is the error rate at which the master frequency Π_ν vanishes [1,2]. The problem with this definition is that, even in the deterministic limit $N \to \infty$, Π_ν never vanishes for *finite* ν. The vanishing of the master frequency is an artifact of neglecting reverse mutations [1], which can be justified in the limit $\nu \to \infty$ only. In particular, for the single-sharp-peak replication landscape this prescription yields a very simple equation for the replication accuracy at the threshold in the deterministic regime [1,2],

$$-\ln q_t = \frac{1}{\nu}\ln a. \tag{17}$$

We must emphasize that for finite ν this equation is an approximation only. A more appropriate definition of the error threshold, which is useful for the finite N case as well, is obtained by considering the statistical properties of the entire molecular population [19]. In particular, we focus on the normalized mean Hamming distance d between the master sequence and the entire population and define the error threshold as the error rate at which the standard deviation σ is maximal [19].

In Figs. 1 and 2 we show the theoretical predictions for d and σ using the steady-state solution of the recursion equations (15). As expected, the effects of the fluctuations in $\Pi_P(t)$ for the different runs are stronger for small N and hence our analytical approximation yields very poor results in this case, although it reproduces quite well the qualitative behavior pattern of the quantities measured. However, already for $N = 100$ there is a good agreement between the theoretical predictions for d and the simulation results, provided that $1 - q$ is not too near the threshold transition. Rather surprisingly, that agreement is better for the standard genetic algorithm. As expected, however, the theoretical predictions for the standard deviation σ are very poor since our approximation scheme neglects the fluctuations in Π_P between the different runs, which are directly measured by σ. Nevertheless, the qualitative features of the simulation results are again well described by the theoretical curves. We note, in particular, the abrupt increase of σ as the error threshold is approached from below and the slow decay as the error rate increases further. The agreement between theory and simulation becomes better as N increases. In Fig. 3 we show the replication accuracy at the threshold q_t as a function of the reciprocal of the population size. The increase of q_t with decreasing N is expected since the fluctuations become stronger for small N and so the replication must be more accurate in order to keep the master string in the population. In particular, for large N we find that q_t increases linearly with $1/N$. This result is in disagreement with the predictions of the birth and death model of error threshold proposed by Nowak and Schuster, which predicts that q_t increases with $1/\sqrt{N}$ for large N [14]. We note that, despite the claim of those authors, it is not possible to discern whether q_t increases with $1/N$ or $1/\sqrt{N}$ from their numerical data obtained using Gillespie's algorithm [14].

Another interesting phenomenon, termed stochastic escape, is the loss of the master string in a finite population [20–22]. In the limit $\nu \to \infty$ this loss becomes irreversible, since no reverse mutation will be able to restore the master string. We can easily derive a lower bound to the probability that the master string is absent from the population using the inequality

$$1 - \bar{n}_\nu \leq Pr\{n_\nu = 0\}, \tag{18}$$

which follows trivially from the fact that $n_\nu \geq 0$. Using $\bar{n}_\nu = N\bar{\Pi}_\nu$, we can find the replication accuracy q_l such that the condition $\bar{\Pi}_\nu = 1/N$ is satisfied for fixed ν and a. Clearly, for $q < q_l$ the probability that the master string is absent from the population is nonzero. However, since this probability may be nonzero for $q > q_l$ as well, q_l gives only a lower bound to the value of the replication accuracy below which the stochastic escape phenomenon actually takes place. This bound is presented in Fig. 3 as a function of the reciprocal of the population size. In the limit $N \to \infty$ we find $q_l \to 0$ for finite ν, since in the deterministic regime $\bar{\Pi}_\nu$ is bounded by $1/2^\nu > 0$. It is interesting that for N not too large we find $q_l > q_t$ so that the master string is likely to be absent from the quasispecies for replication accuracies in the range $q_t < q < q_l$.

We turn now to the analysis of the deterministic regime, $N \to \infty$. In this case, the sum in Eq. (15) is dominated by the closest integer to $(N-1)\overline{\Pi}_\nu(t)$, so that $r \to \overline{\Pi}_\nu(t)$ and the recursion equations (15) reduce to

$$\overline{\Pi}_P(t+1) = \frac{\sum_{R=0}^{\nu-1} M_{PR}\overline{\Pi}_R(t) + aM_{P\nu}\overline{\Pi}_\nu(t)}{1 + \overline{\Pi}_\nu(t)\,(a-1)}, \tag{19}$$

for $P = 0, \ldots, L$.

To appreciate the relevance of the present formulation of the quasispecies model, we compare it with the exact solution of the kinetic equations (1) for finite ν. In fact, as pointed out by Swetina and Schuster [23], for the type of replication landscape considered in this paper, the 2^ν molecular concentrations x_i can also be grouped into $\nu + 1$ distinct classes according to the number of monomers 1 that compose the molecules. This procedure allows the description of the chemical kinetics by only $\nu + 1$ coupled first-order differential equations. In particular, for the single-sharp peak landscape the concentrations of molecules in class $P = 0, \ldots, \nu$, denoted by Y_P with $\sum_P Y_P = 1$, obey the differential equations [23]

$$\frac{dY_P}{dt} = \sum_{R=0}^{\nu-1} M_{PR}\, Y_R + aY_\nu M_{P\nu} - Y_\nu \left[1 + Y_\nu\,(a-1)\right]. \tag{20}$$

It is clear then that both models (19) and (20) possess the same stationary state. This very interesting finding indicates that in the quasispecies model there is no linkage disequilibrium at the population level in the stationary regime, i.e., the population is a random combination of the constituent molecules. In fact, this results holds true for any choice of the replication landscape A_P, as can be easily verified by taking the limit $N \to \infty$ in Eq. (14).

It is also interesting to compare the deterministic limit of our model (19) with the population genetics approach to the deterministic quasispecies model [7]. This can easily be done by writing down a recursion equation for the frequency of monomers 1 in the population $\bar{p}_t = \frac{1}{\nu} \sum_P P \, \overline{\Pi}_P(t)$, namely,

$$\bar{p}_{t+1} = 1 - q + (2q - 1) \, \frac{\bar{p}_t + (a-1) \, \overline{\Pi}_\nu(t)}{1 + (a-1) \, \overline{\Pi}_\nu(t)}. \tag{21}$$

Clearly, this equation is useless since one must solve the general recursion equations (19) in order to find $\overline{\Pi}_\nu(t)$. However, the population genetics formulation makes use of the binomial assumption (4) to set $\overline{\Pi}_\nu(t) = \bar{p}_t^\nu$ so that the recursion equation (21) will involve the monomer frequencies only [7].

In Fig. 4 we present the logarithm of the replication accuracy at the threshold $\ln q_t$ as a function of the logarithm of the selective advantage $\ln a$ for several values of ν for the deterministic case. There is a good agreement with (17) for $a \approx 1$ and, as expected, this agreement becomes better as ν increases. We have also verified that the prediction of the population genetics approach [7] yields a very poor approximation for the location of the error threshold. Furthermore, we have verified that $\overline{\Pi}_P$ departs significantly from a binomial distribution only near the threshold transition. Aside this region, the population genetics approach provides a reliable and concise description of the deterministic quasispecies model.

IV. CONCLUSION

In this paper we have proposed a simple analytical model, based on the neglect of linkage disequilibrium, to study the error propagation in the quasispecies evolution of finite populations. In particular, our finding that in the deterministic regime this model yields exactly

the same stationary state of the original kinetics model [23] implies that the steady-state molecular population of the deterministic quasispecies model is a random assembly of the component molecules.

Some comments regarding the comparison of our approach with previous population theoretical analyses of the finite N quasispecies model [21,24] are in order. These works provide approximate formalisms to study the evolution of finite populations on a multiplicative single-peak fitness landscape without neglecting linkage disequilibrium. Interestingly, for this fitness landscape, which is given by $A_P = (1-\hat{\sigma})^{\nu-P}$ with $0 < \hat{\sigma} < 1$, the binomial assumption (4) yields the exact solution to the deterministic equations for Π_P [21]. Moreover, in the weak selection limit ($\hat{\sigma} \ll 1$) the $1/N$ corrections to the deterministic value of the mean Hamming distance between the master sequence and the whole population can be calculated analytically [21]. An alternative formalism concentrates on the evolution of the ensemble average of the first cumulants of the distribution of fitness in the population [24]. It is not clear, however, whether these quantities can be related to the more natural measures of the population composition, namely, d and σ. We note that the location of the error threshold is not addressed in these works.

An important open question, which can be answered through intensive numerical simulations only, is the dependence of the replication accuracy at the threshold on the population size for large populations. In fact, the numerical data existent in the literature do not allow to distinguish between the $1/N$ dependence predicted by our model and the $1/\sqrt{N}$ dependence predicted by the birth and dead model [14]. We think that simulations based on genetic algorithms rather than on Gillespie's algorithm may prove more effective to address this issue.

ACKNOWLEDGMENTS

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

[1] M. Eigen, Naturwissenchaften **58**, 465 (1971).

[2] M. Eigen, J. McCaskill and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).

[3] I. Leuthäusser, J. Chem. Phys. **84**, 1884 (1986); J. Stat. Phys. **48**, 343 (1987).

[4] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).

[5] S. Galluccio, R. Graber and Y-C. Zhang, J. Phys. A **29**, L249 (1996).

[6] S. Galluccio, Phys. Rev. E **56**, 4526 (1997).

[7] D. Alves and J. F. Fontanari, Phys. Rev. E **54**, 4048 (1996); J. Phys. A **30**, 2601 (1997).

[8] M. W. Feldman, in *Lectures in the Science of Complexity*, edited by D. Stein (Addison-Wesley, Longman, 1989), pp. 501-526.

[9] W. Ebeling and R. Feistel, Ann. Physik **34**, 81 (1977).

[10] J. S. McCaskill, Biol. Cybern. **50**, 63 (1984).

[11] H. Inagaki, Bull. Math. Biol. **44**, 17 (1982).

[12] Y-C. Zhang, Phys. Rev. E **55**, R3817 (1997).

[13] D. Gillespie, J. Comp. Phys. **22**, 403 (1976); J. Phys. Chem. **81**, 2340 (1977).

[14] M. Nowak and P. Schuster, J. Theor. Biol. **137**, 375 (1989).

[15] W. Gabriel, M. Lynch and R. Bürger, Evolution **47**, 1744 (1993).

[16] E. van Nimwegen, J. P. Crutchfield and M. Mitchell, Phys. Lett. A **229**, 144 (1997).

[17] J. Maynard Smith and E. Szathmáry, *The Major Transitions in Evolution* (Freeman, Oxford, 1995).

[18] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, 1989).

[19] T. Wiehe, Genet. Res. Camb. **69**, 127 (1997).

[20] P. G. Higgs and G. Woodcock, J. Math. Biol. **33**, 677 (1995).

[21] G. Woodcock and P. G. Higgs, J. Theor. Biol. **179**, 61 (1996).

[22] L. Peliti, preprint cond-mat/9712027.

[23] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).

[24] A. Prügel-Bennett, J. Theor. Biol. **185**, 81 (1997).

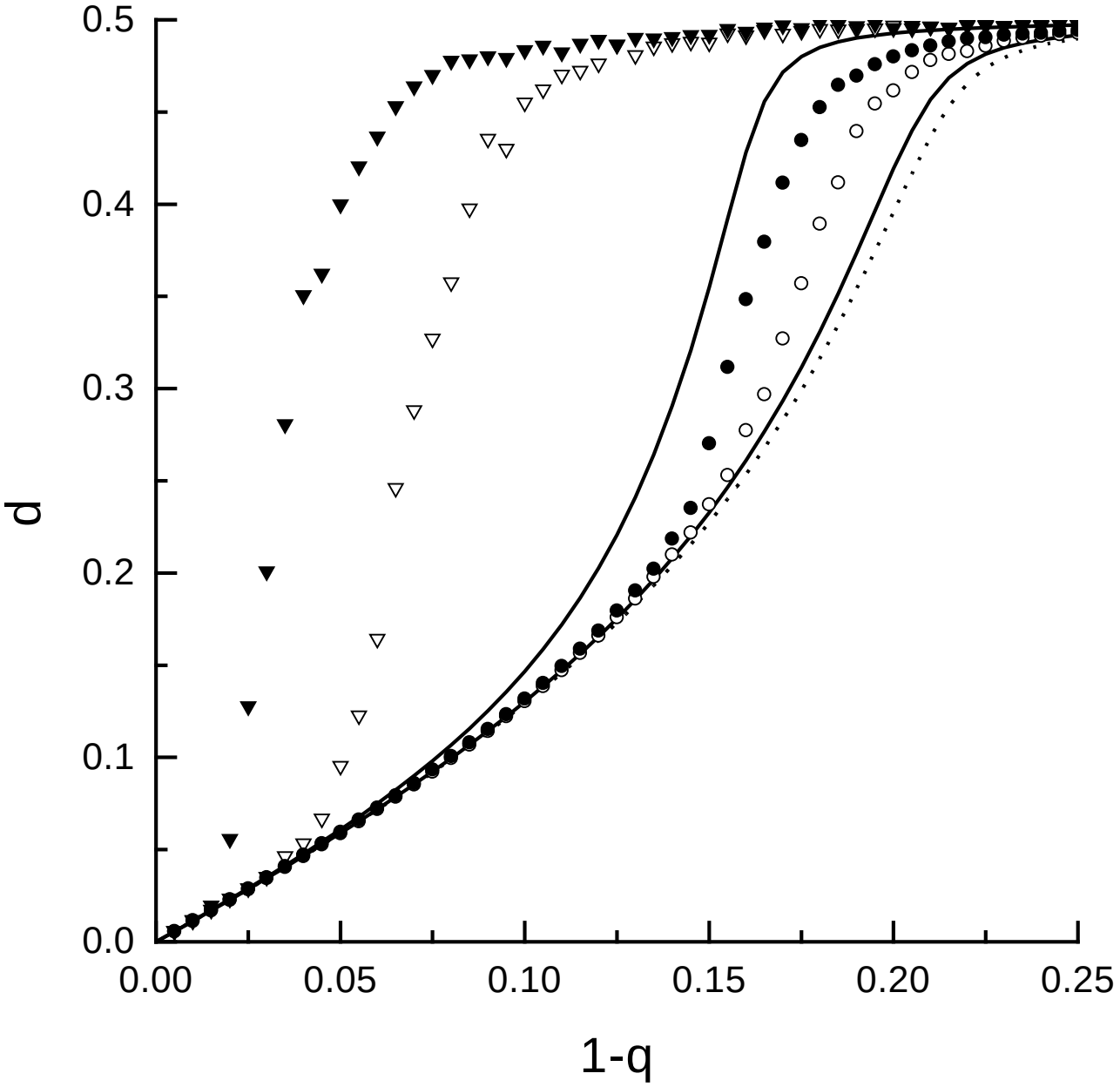
FIGURES

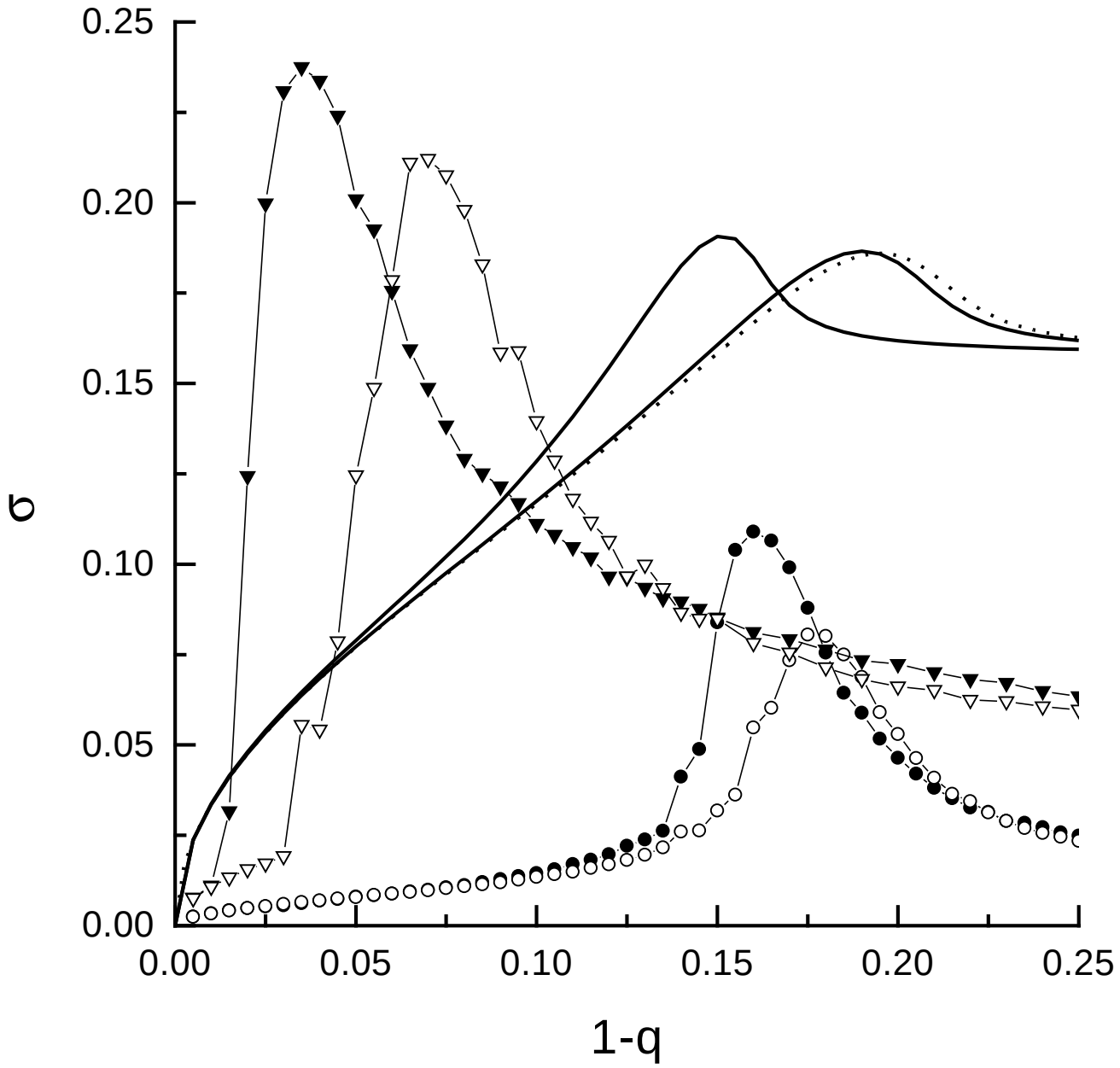
FIG. 1. Steady-state normalized mean Hamming distance between the master sequence and the whole population as a function of the error rate per digit for $N = 10$ ($\bigtriangledown$), and $N = 100$ ($\bigcirc$). The full symbols are the results obtained with the algorithm that neglects linkage disequilibrium, while the empty symbols are the results obtained with the standard genetic algorithm. The theoretical prediction is given by the solid curves. The dashed line is the prediction for $N \to \infty$. The parameters are $\nu = 10$ and $a = 10$.

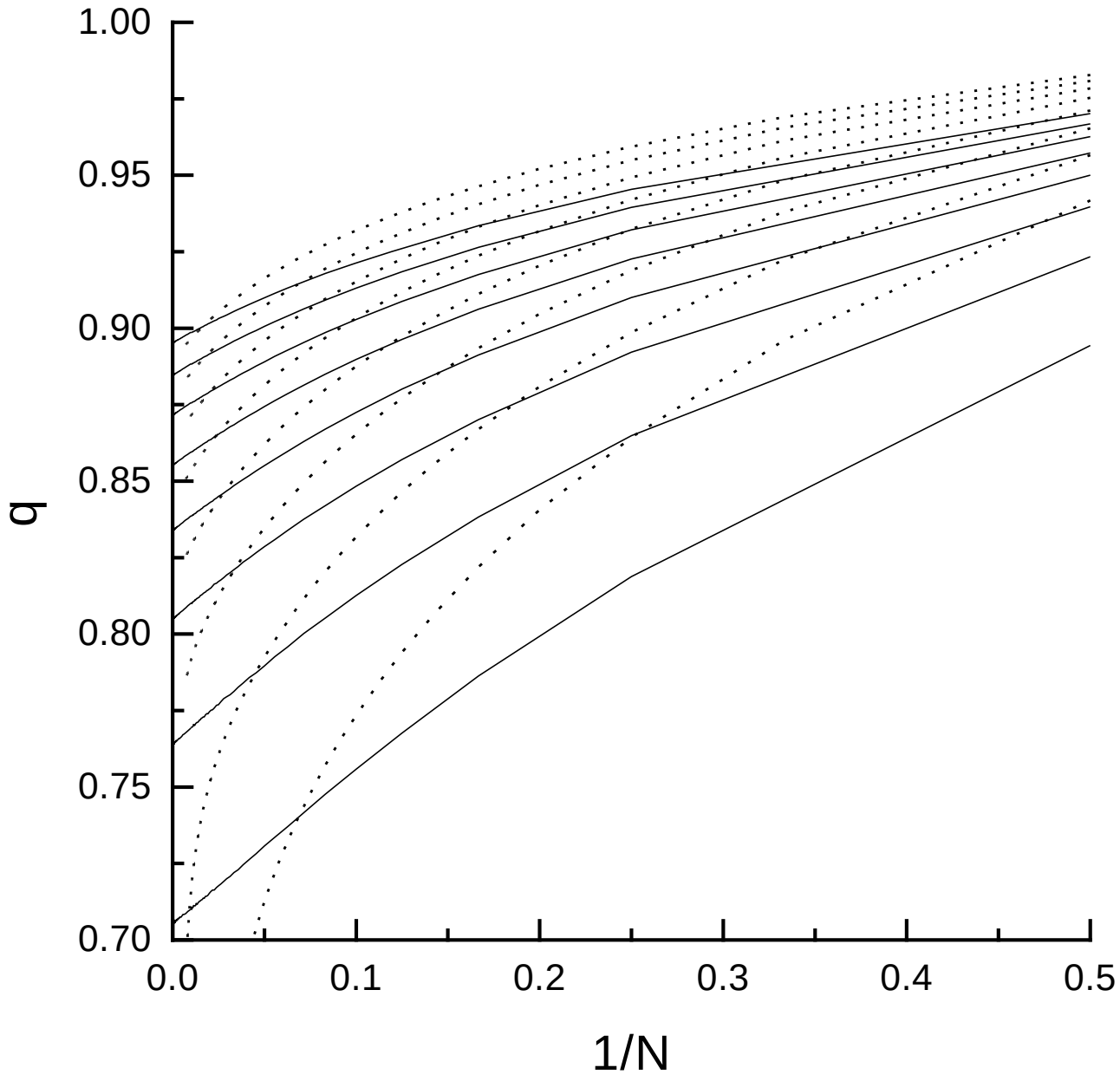
FIG. 2. Steady-state standard deviation of the normalized mean Hamming distance between the master sequence and the whole population as a function of the error rate per digit. The parameters and convention are the same as for Fig. (1).

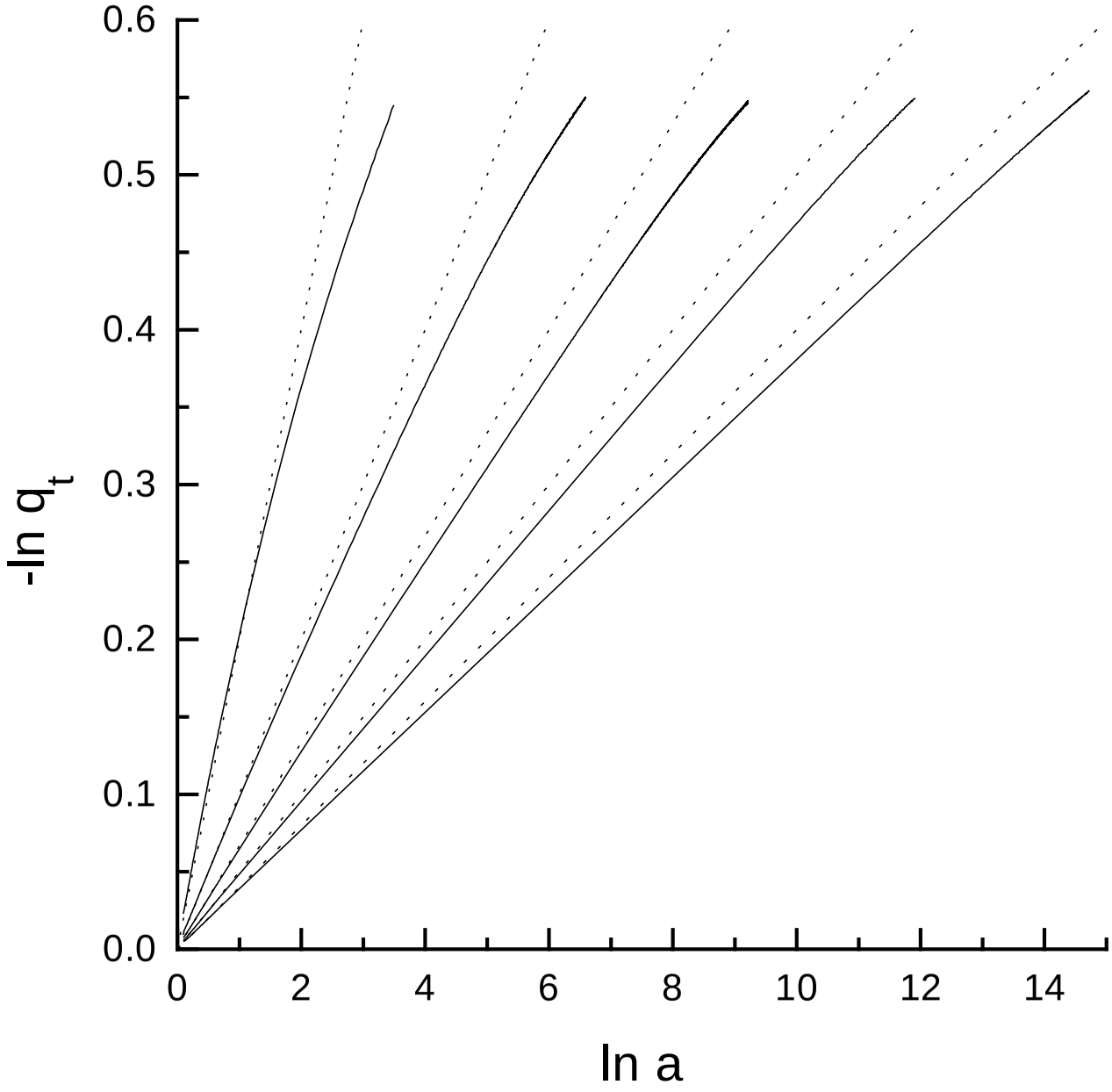
FIG. 3. Replication accuracy at the error threshold q_t (solid curves) and lower bound to the replication accuracy below which the stochastic escape phenomenon occurs q_l (dashed curves) as functions of the reciprocal of the population size for $a = 10$ and (from bottom to top) $\nu = 6, 8, \ldots, 20$.

FIG. 4. Logarithm of the replication accuracy at the threshold in the deterministic regime as a function of the logarithm of the selective advantage for (from top to bottom) $\nu = 5, 10, 15, 20$ and 25. The solid curves are obtained using Eq. (19) and the dashed straight lines are given by Eq. (17).


0.5
0.4
0.3
0.2
0.1
0.0
d
0.00
0.05
0.10
0.15
0.20
0.25
1-q


0.25
0.20
0.15
0.10
0.05
0.00
σ
0.00
0.05
0.10
0.15
0.20
0.25
1-q


1.00
0.95
0.90
0.85
0.80
0.75
0.70
q
0.0
0.1
0.2
0.3
0.4
0.5
1/N


0.6
0.5
0.4
0.3
0.2
0.1
0.0
-ln q_t
0
2
4
6
8
10
12
14
ln a